arXiv:2601.01872v1 [cs.RO] 5 Jan 2026

C AU SA L NAV：面向动态户外场景中自主移动机器人的长期具身导航系统

段宏博[1], 罗尚义[1],邓志远[1],陈彦博[1],蒋元浩[1],刘毅[1], 刘芳明[2],王雪倩[1]

摘要——由于语义推理、动态环境以及长期稳定性方面的挑战，在大型室外环境中实现自主语言引导导航仍是移动机器人学的一项关键难题。我们提出 CausalNav，这是首个专为动态户外环境设计的基于场景图的语义导航框架。我们利用大语言模型构建了多层级语义场景图，称为具身图，以分层方式整合粗粒度地图数据与细粒度对象实体。构建的图谱作为检索增强生成（RAG）的可检索知识库，使系统能够基于开放词汇查询实现语义导航和长程规划。通过融合实时感知与离线地图数据，具身图支持在动态户外环境中跨不同空间粒度实现稳健导航。在场景图构建和分层规划模块中，系统均对动态物体进行显式处理。具身图在时间窗口内持续更新，以反映环境变化并支持实时语义导航。在仿真和真实场景中开展的大量实验表明，该方法具有出色的稳健性和效率。

索引词—语义场景理解、自主车辆导航、人工智能赋能的机器人学

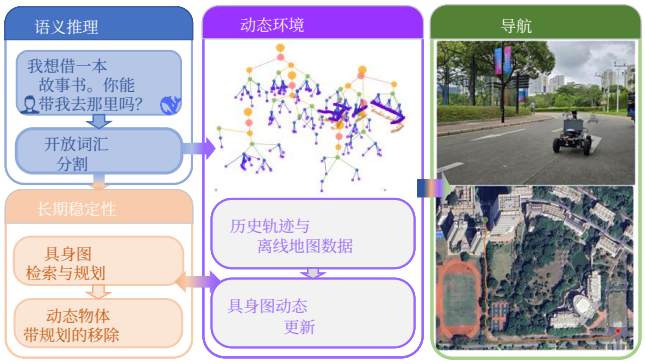

图 1. CausalNav 的整体工作流程。CausalNav 提出了一种新颖的具身导航框架，融合了开放词汇语义推理、动态环境适应和基于具身图的规划。通过利用历史数据和实时数据动态更新导航图，该系统能够在复杂户外环境中实现稳健、长时程、由语言指引的导航。

一、引言

RECENT 移动机器人学的进步，使研究重点超越了传统的控制、感知与导航。终身导航不仅要求准确性，还要求具备语义理解能力、在动态环境中进行增量学习的能力、长期鲁棒性以及决策能力 [1]。这需要进一步推动机器人学与人工智能的深度融合，从而实现可扩展且自适应的自主运行。

由于大规模室外环境具有动态且难以预测的特性，自主导航仍面临挑战 [2], [3]。机器人必须在推理语义和高层任务意图的同时进行局部和全局规划。然而，大多数视觉语言导航基准测试都局限于静态室内环境，并采用逐步指令，这与真实世界场景相去甚远：在真实世界场景中，人们会提出抽象目标，并期望机器人理解语义、推断空间关系，以及从任意起点进行导航 [4]。

传统户外导航方法主要采用点到点导航，并严重依赖高精度地图构建[5]。然而，该领域现有的大多数研究对开放世界环境 [6], [7], 的语义理解和推理能力有限或较为浅层，难以实现持续的人机交互。尽管基于学习的导航方法发展迅速，但其在远距离真实场景中的鲁棒性尚未得到充分验证 [8], [9]。这类方法也难以应对动态环境，且通常需要大规模训练数据。大多数视觉语言导航（VLN）研究仍在室内环境中开展 [10]–[13]，处理环境变化的能力有限。相比之下，室外环境范围广阔、动态性强且往往条件恶劣，对具身导航提出了更大的挑战，并要求算法具备更高的鲁棒性。

为应对上述挑战，我们提出了 CausalNav，这是一种利用离线地图数据增强的开放世界语义导航系统。CausalNav 利用经 RAG 检索增强的、由大语言模型构建的场景图，并结合分层

稿件收到日期：2025年6月11日；修订日期：2025年11月6日；录用日期2026年1月2日。
经副主编和审稿人评议，编辑 Ashis Banerjee 推荐本文发表。本工作得到国家自然科学基金委员会项目（批准号：62293545 和 U21B6002）的资助，部分得到鹏城实验室重大重点项目（批准号：PCL2024A06 和 PCL2025A10）的资助，并部分得到深圳市科技计划项目（批准号：RCJC20231211085918010）的资助。（通信作者：王雪倩。）[1] 清华大学深圳国际研究生院人工智能与机器人研究中心，中国深圳 518055。dhb24@ mails.tsinghua.edu.cn; wang.xq@sz.tsinghua.edu.cn[2] 鹏城实验室，中国 518108。fangminghk@gmail.com数字对象标识符（DOI）：见本页顶部。

全局-局部规划，实现动态户外环境中的开放词汇长距离导航。我们的框架拓展了近期将 RAG 应用于机器人学的研究成果 [14]–[16]，，并引入了关键改进。通过利用城市地图数据和实时感知，我们构建了一个多层级具身图，将城市校园环境中的粗粒度建筑物与细粒度对象实体整合在一起。该图会动态更新，并支持对导航目标进行多尺度记忆表示，同时将时间窗口内的动态物体纳入其中。此外，利用大语言模型进行层次化摘要，通过将场景拓扑组织成语义结构化的层级，提升了具身图的空间抽象能力。

如图1所示，我们的导航框架融合了开放词汇语义推理、动态环境适应和基于具身图的规划，从而实现由语言指引的长程稳健导航，适用于动态户外场景。值得注意的是，我们证明，在边缘端部署大语言模型，无需依赖商业 API （如 GPT-4o ），即可在城市校园环境中实现稳健性能。在自主平台上本地运行的开源模型是一种实用且有效的替代方案。总而言之，我们的主要贡献如下：

1）我们利用大语言模型构建了多层次具身图，整合了从粗粒度地图数据到细粒度物体实体的层次化语义，为语义导航提供可供检索增强生成检索的知识库。2）我们通过在场景图和规划层面持续更新具身图与体素地图，解决动态物体处理问题。该图在时间窗口内动态更新，以支持跨不同空间尺度的稳健检索。3）我们提出 *CausalNav*，这是首个面向动态户外环境的基于场景图的语义导航框架。我们在仿真和真实校园场景中验证了该系统。与最先进的方法相比，CausalNav在导航成功率、轨迹效率和稳健性方面表现更优。

II. 相关工作

A. 传统户外导航方法

在复杂户外环境中实现地面机器人的可靠、高效导航，仍是一项长期挑战 [2]。传统流程将导航分解为彼此独立的模块——感知 [4], 、定位 [17]、路径规划和控制——具有可解释性和鲁棒性。然而，这些流程依赖高精地图（HD）和人工标定，因此在大范围、动态环境中的可扩展性有限。虽然这些方法适用于缺乏大规模数据或端到端训练 [5], 的结构化场景，但它们对开放式任务的泛化能力不足，也无法处理自然语言交互。此外，它们依赖静态地图和人工设计规则，进一步限制了其对真实世界动态场景的适应能力。

B. 基于学习的导航方法

近年来，基于学习的导航，尤其是强化学习（RL）取得了进展，使得能够直接根据视觉输入学习策略 [18]。RECON [6] 和ViKiNG [7] 等方法结合潜在目标模型、拓扑记忆和可通行性估计，以支持探索和局部规划。与此同时，扩散模型已应用于局部轨迹生成，NoMaD [8]和 ViNT [19] 等方法采用基于扩散的动作或目标采样，以实现高效规划和避障 [20]。尽管这些方法颇具创新性，但通常受限于短距离导航，缺乏明确的全局路径推理能力。此外，它们往往需要大量特定环境的训练数据，并且受仿真到现实差距影响，泛化能力有限。这些方法在动态或长时域真实世界场景中的鲁棒性仍未得到充分验证。

C. 基于场景图的导航方法

基于场景图的视觉语言导航发展迅速，尤其是在室内环境中 [10]–[13]。场景图提供了高层次的空间抽象，使感知与传感器解耦，并支持层次推理。Hydra [10]构建了从度量层到语义层的分层三维场景图，而 VLMaps [12] 则将预训练的视觉语言特征与三维地图融合。LExis [13] 将视觉与语言表征对齐，用于语义同步定位与建图和拓扑规划。然而，这些方法大多仍局限于室内环境，而室外尺度和动态性带来了重大挑战。

随着基础模型的兴起，近期研究开始将大语言模型整合到语言条件规划中。SayPlan [11] 将大语言模型与三维场景图相结合，以实现可扩展的任务推理；LM-Nav [21] 则支持针对新物体目标的指令跟随。然而，这些系统通常将大语言模型或视觉语言模型视为独立的评估器，而非集成建图框架的组成部分。此外，幻觉和遗忘限制了它们在动态实时室外场景中的可靠性。

III. 方法

A. 系统概述

如图2所示，CausalNav 采用由感知、图构建和规划组件组成的层次化架构，专为动态户外环境中的长时域导航而设计。受 OpenGraph[22], 启发，系统构建了多层级具身图 G，由物体节点 ν_i^{obj}、自车节点 ν_i^l、建筑节点 ν_i^{build}和层次聚类节点 $\nu_i^{cluster}$组成。该图融合了粗粒度地图信息与细粒度语义实体。其详细结构如图4所示。在第 III-B节中，系统利用开放词汇感知模型 [23] 和基于激光雷达的定位 [24], 实现语义对齐，从而提取物体节点和自车节点信息。

在第 III-C 节中，系统通过多目标跟踪区分动态、静态和准静态物体。为减轻环境动态性对构建

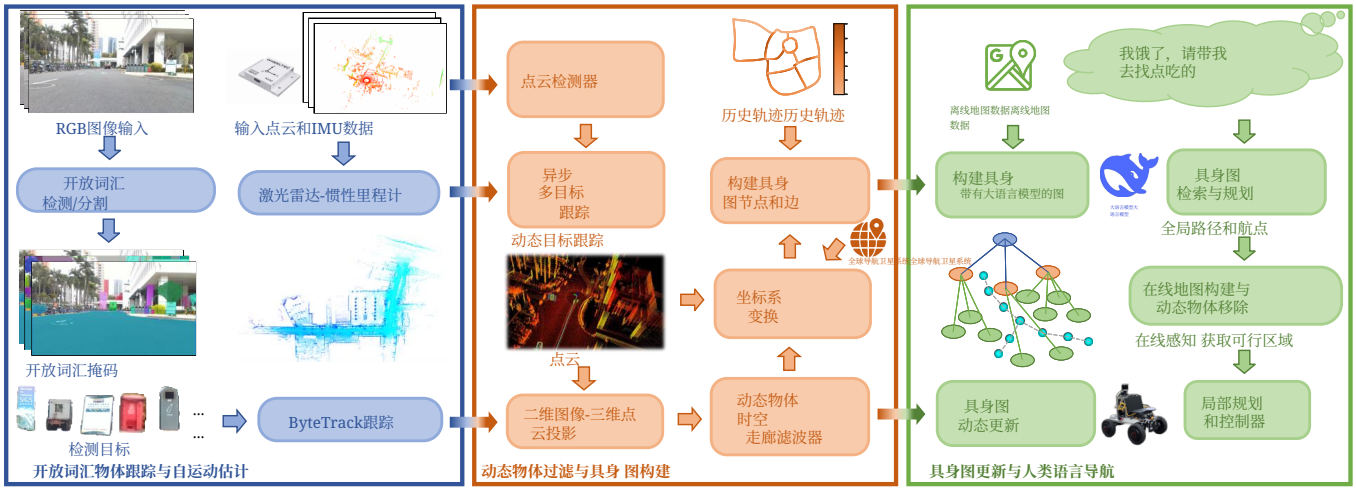

图2。CausalNav框架由三个依次运行的模块组成：(1) 开放词汇物体跟踪与自运动估计（第III-B节）：融合RGB、激光雷达和惯性测量单元的输入，进行开放词汇目标检测与跟踪，并通过二维—三维时空对齐估计自运动。(2) 动态物体过滤与具身图构建（第III-C节）：通过异步多目标跟踪过滤瞬态动态物体，并构建时间上稳定的具身场景图。(3) 具身图更新与人类语言导航（第III-D节）：利用大语言模型构建并更新语义图，以解析语言指令并进行分层规划，同时实时移除动态物体，从而实现稳健导航。

具身图的影响，在更新语义图之前会应用时空过滤机制。通过引入建筑节点，并借助大语言模型（LLM）进行层次聚类，所构建的具身图支持检索增强生成，使系统能够有效响应人类语言指令。

在第 III-D 节中，系统持续融合离线地图与实时感知数据，以实时更新图。全局导航依赖预存路线或动态生成的场景图，而局部路径规划则通过实时激光雷达建图并即时移除动态障碍物来进行自适应规划。这种设计实现了灵活的语言驱动导航，同时具备长期稳定性和语义感知能力。

B. 开放词汇目标跟踪与自运动估计

1）开放词汇目标跟踪：为了获取有关物体节点描述、形状、大小和位置的丰富信息，我们将物体状态表示为 ${}^{w}\mathbf{T}_{\text{obj}} = \{{}^{w}\mathbf{R}_{\text{obj}},\ {}^{w}\mathbf{p}_{\text{obj}}\}$，其中物体位姿 ${}^{w}\mathbf{T}_{obj}$ 表示在 $SE(3)$ 空间中。在室外环境中，我们使用 YOLO-World [23]，这一轻量级开放词汇检测器，从每个 RGB 帧 I_t 中提取二维边界框和分割掩码。随后，由 ByteTrack [25]，实现的多目标跟踪函数 $\mathcal{C}(\cdot)$ 对这些检测结果进行时序关联，从而获得稳定的跟踪结果：

$$\mathcal{S}_t = \mathcal{C}(\text{YOLO-World}(I_t)). \tag{1}$$

这里，$\mathcal{S}_t$ 表示当前帧中检测到的物体集合，定义为

$$\mathcal{S}_t = \{S_i = (c_i, \text{2DBBox}_i, \mathcal{B}_i) \mid i = 1, 2, \ldots, n\}, \tag{2}$$

其中，$\mathcal{B}_i$ 是分割掩码，2DBBox_i 是二维边界框，c_i 是第 i 个跟踪物体的描述。

由于深度相机在真实户外环境中的性能有限，我们融合相机和激光雷达数据进行物体定位。原始激光雷达点云表示为 $\mathcal{P} = \{\mathbf{P}_i = (x_i, y_i, z_i)\}_{i=1}^{N}$。对于当前时间步 t，每个激光雷达点 $\mathbf{P}_i \in \mathcal{P}_t$ 通过标定相机投影模型投影到帧 I_t 的图像平面上，得到图像空间坐标 ${}^{c}\mathbf{p}_i$，计算如下：${}^{c}\mathbf{p}_i = \mathbf{K} \cdot \mathbf{H} \cdot \mathbf{P}_i$，其中$\mathbf{K}$ 是相机内参矩阵，$\mathbf{H}$ 是通过激光雷达-相机联合标定 [26]得到的外参变换矩阵。

随后，我们定义相机坐标系下物体特定的三维点云：

$${}^{l}\mathcal{P}_{\text{obj}} = \{\mathbf{P}_i \in \mathcal{P}_t \mid {}^{c}\mathbf{p}_i \in \mathcal{B}_i\}. \tag{3}$$

根据筛选后的集合 ${}^{l}\mathcal{P}_{\text{obj}}$，构建一个最小体积三维边界框。将该边界框的质心作为物体的估计三维位置，记为 ${}^{l}\mathbf{p}^{\text{obj}}$。

我们从激光雷达-惯性里程计系统 [24] 获取当前自车位姿 ${}^{w}\mathbf{T}_{l}$。检测到的物体在激光雷达坐标系下的相对位姿记为 ${}^{l}\mathbf{T}^{\text{obj}}$。对应的世界坐标系下的物体位姿由 ${}^{w}\mathbf{T}_{\text{obj}} = {}^{w}\mathbf{T}_{l} \cdot {}^{l}\mathbf{T}^{\text{obj}}$ 计算得到。为增量更新拓扑图，每个跟踪物体都会贡献一个节点，定义如下：

$$\nu_i^{obj} = \{c_i, \text{3DBBox}_i, {}^{w}\mathbf{p}_{\text{obj}}\}, \tag{4}$$

并将其插入或更新到具身图 G中。

对于每个跟踪物体 ν_i^{obj}，如果它不在图 G中，则创建一个新节点；否则，更新其位置：

$$G \leftarrow \begin{cases} G \cup \{\nu_i^{obj}\}, & \text{if } \nu_i^{obj} \notin G \\ G \setminus \{{}^{old}\nu_i^{obj}\} \cup \{\nu_i^{obj}\}, & \text{if } {}^{old}\nu_i^{obj} \in G. \end{cases} \tag{5}$$

2）自运动估计：对于自车，其状态表示为 ${}^{w}\mathcal{E}_l = [{}^{w}\mathbf{R}_l,\ {}^{w}\mathbf{p}_l,\ {}^{w}\boldsymbol{\xi}_l]$，其中 ${}^{w}\mathbf{R}_l$ 表示方向，${}^{w}\mathbf{p}_l$ 表示位置，${}^{w}\boldsymbol{\xi}_l$ 表示世界坐标系中的速度

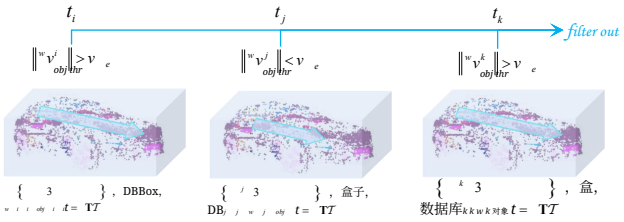

图3. 时空走廊内三个观测轨迹点及其对应三维边界框的示意图。点云展示了在不同时间戳采集的同一车辆。

当运动步长超过阈值d时，位置和速度将作为自车节点 $\nu_i^l = \{{}^w\mathbf{p}_l,\ {}^w\boldsymbol{\xi}_l\}$存储，并向 G添加一条新边 $e_i^l = (\nu_i^l, \nu_{i-1}^l)$。如图4所示，表示历史轨迹的自车节点 ν_i^l 和边 e_i^l 存储在具身图中，随后用于第 III-D2 节所述的全局规划任务。

C. 动态物体过滤与具身图构建

1) 动态目标时空走廊过滤器：为了稳健地处理室外环境中的动态物体，我们采用了一个将基于鸟瞰图的检测与运动估计相结合的多目标跟踪流程。具体而言，CenterPoint [27] 执行实时点云检测，并通过 TensorRT 加速；而 LIOsegmot [28]则通过融合检测结果与激光雷达-惯性里程计来估计目标速度。由此可将目标可靠地分类为动态、静态或准静态。

为减轻传统基于速度的过滤器 [29], 在具身图中造成的假阳性问题，我们利用历史边界框将每个目标的轨迹编码为空间-时间走廊，如图3所示。这种表示方式增强了对动态实体的过滤能力，并保持了图的结构完整性。

$$\mathcal{T} = \left\{{}^w\mathbf{T}_{\text{obj}}^i, \text{3DBBox}_i, t_i\right\}_{i=1}^n. \tag{6}$$

当目标的位移超过 k步时，将排除其时空走廊 $\mathcal{T}$，并从图构建中移除相应的动态节点 ${}^d\nu_i^{\text{obj}}$：

$$G \leftarrow G \setminus \{\mathcal{T} \mid \mathcal{T} \in D\}. \tag{7}$$

这种基于走廊的过滤能有效地从具身图中移除瞬态或移动目标，减少运动诱发的误差。对于具有间歇性运动模式的实体，例如交叉路口附近的车辆，这种方法尤为有效。

2) 具身图的分层构建：除了第III-B2节所述的自车节点外，我们还通过时空走廊过滤瞬态动态实体后，构建静态环境的分层空间语义表示。由此得到的具身图融合了几何信息与语义信息，为下游推理和导航奠定基础。

静态拓扑图包含两种节点类型：(1) 建筑节点 $\nu_i^{\text{build}} = \{c_i^{\text{build}}, {}^w\mathbf{p}_{\text{build}}\}$，从离线地图中提取并赋予最高层级 L；(2) 物体节点 ν_i^{obj}，表示第 $L-1$ 层级上的细粒度元素（例如消防栓、公交车站）。它们共同构成了环境的多层级空间语义抽象。

随后，[15], 使用空间语义相似度量进行层次聚类：

$$\kappa_{ij} = (1-\alpha)\kappa_{ij}^{\text{spatial}} + \alpha\kappa_{ij}^{\text{semantic}}, \tag{8}$$

其中，$\kappa_{ij}^{\text{spatial}} = \exp(-d_{\text{haversine}}(i,j)/\theta)$，且 $\kappa_{ij}^{\text{semantic}}$是嵌入 $\mathbf{e}_i$ 与 $\mathbf{e}_j$之间的余弦相似度。基于嵌入的相似度增强了对大语言模型标签变体（例如，“垃圾桶”的不同称谓）的鲁棒性。

第$L-1$ 层物体节点自底向上分组，形成第L 层簇 $\nu_i^{\text{cluster}} = \{\text{LLM}(r),\ \text{质心}(r)\}$。每个簇$r \in \mathcal{R}$ 包含若干节点，其语义由大语言模型概括，位置则由其成员的均值确定。聚类过程递归进行，同时考虑簇和建筑节点，直至收敛。

3) 在具身图上进行语义检索：给定查询 q，使用基于大语言模型的选择进行分层语义检索。在第 l层，节点 $n_l \in \mathcal{L}_l$ 的选择概率为：

$$\pi(n_l \mid q) = \frac{\exp[\gamma \cdot \text{LLM}(q, C(n_l))]}{\sum_{n' \in \mathcal{L}_l} \exp[\gamma \cdot \text{LLM}(q, C(n'))]}, \tag{9}$$

其中，$C(n_l)$表示节点描述，$\gamma > 0$ 用于控制锐度。

分层路径得分为：

$$\Lambda(\zeta) = \prod_{l=1}^{D} [\pi(n_l \mid q) \cdot \phi(n_l, n_{l-1})], \tag{10}$$

其中，$\phi(n_l,\ n_{l-1}) = \mathbf{1}_{\{n_{l-1} \in \text{子节点}(n_l)\}}$ 用于确保父子链接有效。

如果已知智能体位置 $\mathbb{L}$，则根据混合得分对候选项重新排序：

$$\eta(n) = \beta\kappa^{\text{spatial}}(n, \mathbb{L}) + (1-\beta)\Lambda(\zeta), \tag{11}$$

其中，$\kappa^{\text{spatial}}(n, \mathbb{L}) = \exp[-d_{\text{haversine}}(n, \mathbb{L})/\theta]$ 用于衡量空间邻近度，而 $\beta \in [0,1]$ 则兼顾空间与语义相关性。

在候选路径中评估 $\eta(n)$，可使系统选出与查询在语义和空间上都最为匹配的节点，以供下游规划使用。

D. 具身图更新与人类语言导航

1) 在线具身图更新：本质上，具身图既充当场景操作器，也充当记忆库，将时间、场景、目标和事件等元素抽象为结构化记忆表示。其核心价值在于帮助机器人理解复杂环境——也就是说，实现一种抽象推理，从而在导航任务中进行更智能、更多样化的决策。在具身图的在线构建与更新过程中，我们重点关注以下方面：

算法1 在线具身图更新

要求： $\mathcal{C}$，k

$G \leftarrow \emptyset$，$t \leftarrow 0$

当系统运行**时**

　$t \leftarrow t+1$，$\mathcal{S}_t \leftarrow \mathcal{C}$（$I_t$，$\mathcal{P}_t$，惯性测量单元）

　对所有 $S_i \in \mathcal{S}_t$ **执行**

　　计算 ${}^{w}\mathbf{T}_{\text{obj}}$，$\mathcal{B}_i$，三维边界框$_i$

　　$\nu_i^{obj} \leftarrow \{c_i, 3$三维边界框$_i$，${}^{w}\mathbf{p}_{\text{obj}}\}$

　　如果 $\nu_i^{obj} \notin G$，**则** $G \leftarrow G \cup \{\nu_i^{obj}\}$

　　else G。**更新**（ν_i^{obj}　）

　　结束条件判断

　结束循环

　对于所有 ${}^{d}\nu_i^{obj} \in \{\mathcal{T} \mid \mathcal{T} \in D\}$，**执行**

　　如果 ${}^{d}\nu_i^{obj}.$**步骤** $\geq k$，**则**

　　$G \leftarrow G \setminus \{\mathcal{T}\}$

　　结束条件判断

　结束循环

　更新 ν_i^l, E_ν 对所有 $\nu \in G$

　$\mathcal{R} \leftarrow$ **HCluster**(G)

　对所有 $r \in \mathcal{R}$ **执行**

　　$E_r \leftarrow \{(\nu_i^{cluster}, \nu_i^{obj}) \mid \nu_i^{obj} \in r\}$

　　$G \leftarrow G \cup \{\nu_i^{cluster}\} \cup E_r$

　循环结束

当条件不成立时结束循环

返回 G

- 场景的多粒度拓扑表示；
- 单个场景内静态物体之间的空间关系；
- 在时间窗口内对动态物体进行建模与管理；
- 场景间关系的隐式编码。

具身图的在线更新过程在算法1中进行了形式化描述。

2) 人类语言导航：CausalNav根据人类语言指令执行导航。基于第III-C3节所述的语义检索方法，系统首先利用实时构建和更新的具身图，从人类语言输入中推断全局目标位置。如果目标无法从机器人的历史轨迹到达，则使用离线路网数据或 Google 地图、高德地图等外部 API 生成粗略全局路线。如果目标与过往轨迹连通，则计算基于 Dijkstra 算法的最短路径。最终生成的全局路径表示为航路点序列$\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n\}$。

在真实世界的室外环境中，机器人必须在高度动态的场景中运行。动态物体留下的残余运动轨迹常常会在地图上表现为静态障碍物，从而降低定位和导航性能。为解决这一问题，我们采用 RH-Map [30], ——一种用于实时局部建图和动态物体移除的三维分区哈希地图框架。

移动机器人通过 RH-Map 可以获取的规划可行区域记为 $\mathcal{F} \subseteq \mathbb{R}^3$。在 $\mathcal{F}$ 内，使用 informed-RRT* 算法生成初始轨迹 [31]：

$$\mathcal{Z}_{\text{init}} = \{\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_N\}, \quad \mathbf{z}_i = [x_i, y_i, z_i]^T. \tag{12}$$

随后，通过 B样条插值对路径 $\mathcal{Z}_{\text{init}}$ 进行平滑。根据相邻航点的方向，为每个插值点 θ_i 赋予一个方向。这便得到一条带方向的参考轨迹：

$$\mathbf{X}_g = \{\mathbf{x}_g^0, \mathbf{x}_g^1, \ldots, \mathbf{x}_g^M\}, \quad \mathbf{x}_g^i = [x_i, y_i, z_i, \theta_i]^T \in \mathbb{R}^4. \tag{13}$$

给定当前机器人状态 $\mathbf{x}_t = [x_t, y_t, z_t, \theta_t]^T$和控制输入 $\mathbf{u}_t = [v_t, \omega_t]^T$，我们将带控制障碍函数的非线性模型预测控制 [32] (NMPC-CBF)问题表述为一个约束优化问题：

$$\min_{\{\mathbf{x}_k, \mathbf{u}_k\}} \sum_{k=0}^{N-1} \left(\left\| \mathbf{x}_k - \mathbf{x}_g^k \right\|_Q^2 + \left\| \mathbf{u}_k \right\|_R^2 \right) \tag{14a}$$

$$\text{s.t.} \quad \mathbf{x}_{k+1} = f(\mathbf{x}_k, \mathbf{u}_k) \tag{14b}$$

$$\mathbf{x}_0 = \mathbf{x}_{\text{init}} \tag{14c}$$

$$\mathbf{x}_k \in \mathcal{X}, \quad \mathbf{u}_k \in \mathcal{U}, \tag{14d}$$

$$\Delta h_{\text{ob}}^i(\mathbf{x}_k, \mathbf{u}_k) + \lambda_k h_{\text{ob}}^i(\mathbf{x}_k) \geq 0. \tag{14e}$$

其中，$\mathcal{X} \subseteq \mathbb{R}^n$ 和 $\mathcal{U} \subseteq \mathbb{R}^m$ 分别表示状态和控制输入的可行集，并纳入了物理约束和安全约束。最后一个约束引入了控制障碍函数，以保证由行人或车辆等动态障碍物定义的安全集具有正向不变性。每个障碍函数 $h_{\text{ob}}^i(\mathbf{x}) > 0$定义了与障碍物 i之间的安全裕度，并随其预测轨迹变化，通常表示为 $h_i(\mathbf{x}) = (x - x_i^p)^2 + (y - y_i^p)^2 - d_{\text{safe}}^2$，其中 $[x_i^p, y_i^p]^T$表示障碍物 i的预测位置，而 d_{safe} 表示安全半径。

IV. 实验与结果

A. 实验设置

仿真环境。如图4所示，我们在城市环境中仿真了一台基于 Gazebo 的地面机器人，该机器人配备了 RealSense D435i 和三维激光雷达。导航过程中，系统会动态构建并可视化具身图。基于 Gazebo 的状态估计器会记录任务完成情况和轨迹。由于缺乏合适的开源替代方案，我们将方法与四种具有代表性的基于学习的系统进行比较：NoMaD、ViNT、GNM 和 CityWalker；这些系统均通过避障实现目标驱动运动预测。为确保公平比较，我们收集图像数据，为 NoMaD、ViNT 和 GNM 构建拓扑地图，并将我们生成的航点用于 CityWalker。我们在随机抽取的 25 项任务上进行评估（每项任务试验 10 次），采用四项指标：成功率（SR）、路径长度加权成功率（SPL）、碰撞次数（CC）和轨迹长度（TL）。机器人到达距目标 10 米以内即视为任务成功。系统运行于 Intel i9-14900K CPU 和单块 RTX 3090 GPU 上，关键模块均可实时运行：开放词汇目标跟踪和自运动为 30 Hz，时空走廊滤波为 20 Hz，局部动态建图和规划为 10 Hz，层次聚类或具身图更新为 1 Hz。

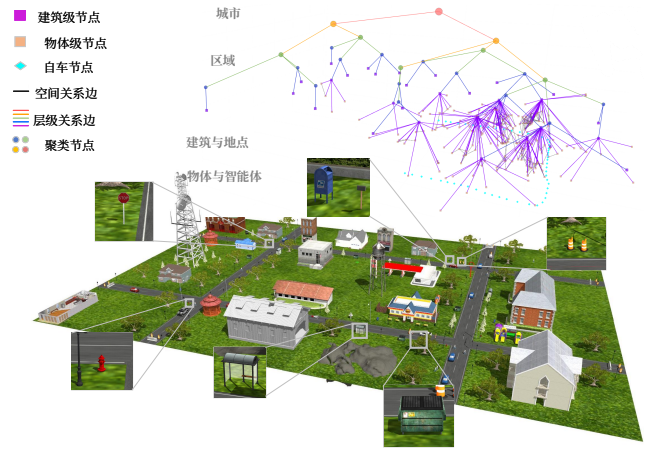

图4. 仿真环境及构建的具身图。该环境包含粗粒度目标（例如建筑物）和细粒度目标（例如消防栓、邮箱）。具身图融合了这两个层级，并随着智能体移动而动态更新。

真实环境。 如图5所示，我们将系统部署在一台轮式移动机器人上，该机器人配备 Intel Core i9-13900H CPU 和 NVIDIA GeForce RTX 4070 GPU。RealSense D435i 相机采集 RGB 图像，RSHelios 激光雷达和 RTK GNSS/INS 模块则以 10 Hz 的频率提供定位数据，精度为 5 cm。我们结合 FAST-LIO2 [24] 与 RTK GNSS/INS 进行坐标变换和精确定位。我们在一个大型校园式户外环境中测试系统，以评估其在动态真实场景中的有效性和稳健性。

B. 仿真环境实验

本实验评估我们的方法在长距离导航任务中的表现，重点关注成功率、导航效率以及对动态环境的适应能力。在仿真中，行人和车辆模拟了现实世界的复杂性。根据导航目标的空间尺度，任务分为短距离、中距离和长距离场景。结果汇总于表 I。从 SR 和 SPL 来看，CityWalker 和 CausalNav 均表现出色。然而，对长距离任务中 TL 的分析表明，ViNT、NoMaD 和 GNM 等拓扑方法受单向连通性的限制，导致路径规划效率低下；即使目的地可以到达，轨迹也会显著变长。尽管 CityWalker 和 CausalNav 在 SR、SPL 和 TL 上表现相当，但二者在碰撞次数（CC）方面存在显著差异。CityWalker 作为一种基于学习的方法，无需大量参数调优便易于部署。然而，它在动态环境中的泛化能力有限，往往无法避开移动障碍物。相比之下，CausalNav 通过实时构建安全区域，并在约 100 毫秒内完成感知和重新规划，增强了对动态环境的响应能力，从而在动态场景中实现更稳健的导航。总的来说，CausalNav 在多项指标上优于现有基线方法，在长距离任务的成功率、轨迹

效率和动态适应能力。

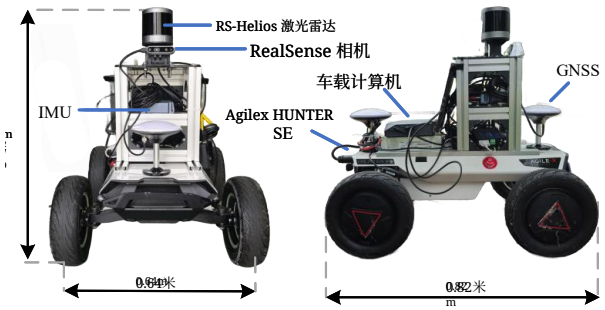

图 5. 真实世界导航实验中使用的机器人。

不同大语言模型在导航中的表现。大语言模型（LLM）在导航的具身图构建和语义推理中发挥着关键作用。然而，由于网络访问受限、延迟高以及数据隐私方面的顾虑，在真实世界部署中依赖在线 API 并不现实。这凸显了在自主导航中本地部署大语言模型的必要性。我们在具身图在线构建的模拟环境中，评估了四种规模相近的近期开源模型，以及在线 API GPT-4o。导航性能总结见表 II。尽管 GPT-4o 的性能最佳，但由于具身图中的分层语义检索提高了查询准确率并减少了幻觉，因此它相较于较小规模开源模型的优势并不显著。在开源模型中，DeepSeek-R1-Distill-14B 的成功率最高，性能也最稳定。

具身图动态更新的影响。在室外环境中，行人和车辆等局部变化可能造成暂时遮挡和可通行性变化，从而破坏具身图与实时感知之间的一致性。尽管初始场景图是通过对整个环境进行探索构建的，但如果不持续更新，它很快就会过时。我们通过一系列目的地随机生成的连续导航任务评估此类更新的影响，以模拟真实世界场景。如表 III 所示，与静态基线方法相比，动态更新策略显著提升了成功率（SR）、成功加权路径长度（SPL）、碰撞次数（CC）和轨迹长度（TL）指标。这些结果表明，即使不存在全局语义变化，维护最新的局部信息对于准确的语义检索和可靠的导航性能也至关重要。

运行效率。为进一步评估计算效率，我们在相同的硬件配置下，将平均单周期延迟与具有代表性的基线方法进行比较。如表 IV 所示，CausalNav 实现了实时性能（10 Hz），且开销仅比 NoMaD 高 11%，展现出出色的运行效率，同时还能提供更丰富的语义推理和动态图维护能力。

语义检索参数消融实验。 如图 6 所示，准确率和召回率呈钟形趋势，在 $\alpha{=}\beta{=}0.5$和 $\gamma{=}1.5$附近达到峰值，这表明空间—语义融合平衡，且大语言模型推理稳定。随着空间权重增大，轨迹长度略有缩短，说明适度的空间偏置能够提升效率，同时不会

表 I 仿真环境中各种语义导航方法的导航性能对比分析。

方法	小型				中型				大型			
	成功率(%)↑	成功加权路径长度(%)↑	CC↓	轨迹长度(m)↓	成功率(%)↑	成功加权路径长度(%)↑	CC↓	轨迹长度(m)↓	成功率(%)↑	成功加权路径长度(%)↑	CC↓	轨迹长度（米）↓
ViNT [19]	84	68.4	0.6	47.16	64	47.8	0.9	92.74	48	32.2	1.6	160.03
NoMaD [8]	82	70.9	0.8	45.61	43	34.2	1.3	**81.13**	22	14.6	2.3	173.72
GNM [33]	84	72.3	0.5	42.02	26	15.2	1.5	88.32	0	0	-	-
CityWalker [9]	**100**	82.4	1.2	43.07	85	73.6	3.4	86.32	**80**	**68.3**	4.5	**136.63**
CausalNav	**100**	**88.9**	**0.2**	**40.66**	**92**	**82.2**	**0.6**	83.16	**80**	66.0	**1.2**	141.82

表 II 不同大语言模型在导航中的性能

方法	成功率(%)↑	成功加权路径长度(%)↑	CC↓	轨迹长度(m)↓
phi4-14B	83	69.5	1.1	106.27
Qwen3-14B	82	70.4	**1.0**	**102.45**
Gemma3-12B	84	67.6	1.2	110.32
DeepSeek-R1-Distill-14B	85	72.1	1.1	103.63
GPT-4o	**88**	**75.3**	**1.0**	103.24

表 III 在线具身图更新对导航的影响

方法	成功率(%)%↑	成功加权路径长度(%)%↑	CC↓	轨迹长度(m)%↓
CausalNav †	78	54.7	1.8	120.35
CausalNav ‡	**90**	**80.1**	**1.1**	**98.25**

†不更新具身图；‡更新具身图。

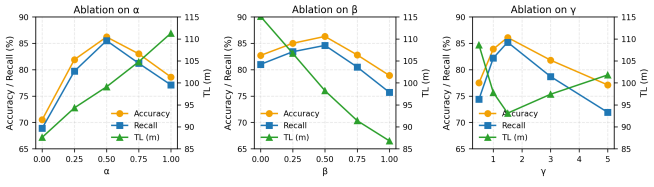

图6. 关键参数的消融结果。从左到右： α、 β和 γ。左轴：准确率 / 召回率（%）；右轴：轨迹长度（m）。

牺牲语义精度。

C. 真实环境实验

我们将 CausalNav 部署在机器人平台上，并与 ViNT、NoMaD、GNM 和 CityWalker 进行了真实场景对比，如图7所示。实验在两种条件下进行。第一种场景（图7(a)）是在对象级指令下进行短距离导航（130 m），只有 ViNT 和 CausalNav 成功完成任务。第二种场景（图7(b)）是在建筑级指令下测试长距离导航（512 m）；只有 CausalNav 完成了任务，其他方法均因碰撞而失败。仿真结果进一步在真实场景中得到验证。ViNT 和 CausalNav 均能成功完成约100米距离的导航任务；然而，只有 CausalNav 能在超过500米的高度动态户外环境中成功执行长距离导航。尤其是，与仿真相比，CityWalker 在真实环境实验中的性能显著较低。我们观察到

表 IV 平均每周期延迟运行时间比较

方法	NoMaD [8]	ViNT [19]	GNM [33]	CityWalker [9]	CausalNav
运行时间（毫秒）↓	**95**	150	110	180	105

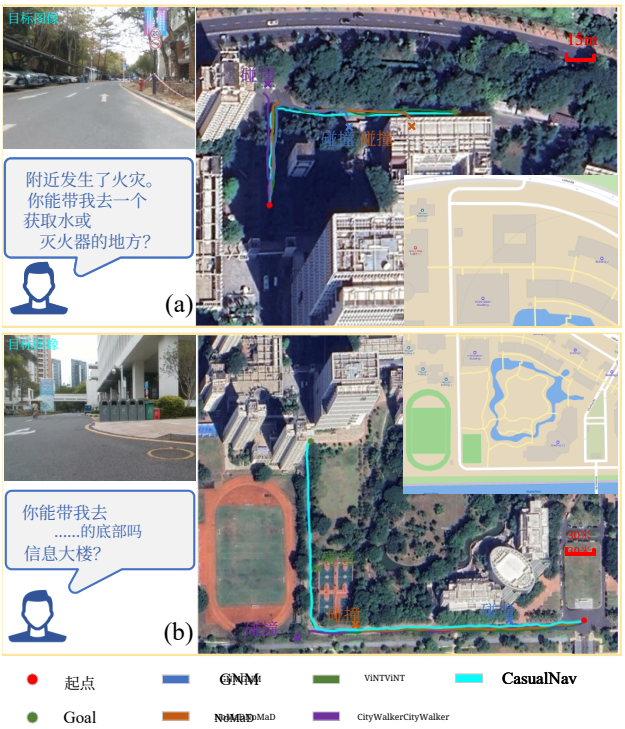

图7. 真实世界场景下不同距离尺度的实验。

CityWalker 对光照条件和环境变化高度敏感，导致轨迹预测不稳定，且对动态物体的处理能力不足。尽管它在仿真中展现出较高的长距离任务成功率，但现实世界会引入仿真中未涵盖的额外复杂性。尤其是，在仿真中可以容忍的轻微碰撞，可能会导致真实世界部署中的任务失败。这些发现证实了 CausalNav 在复杂且动态的城市环境中进行长距离语义导航的鲁棒性。

V. 结论与未来工作

本文介绍了*CausalNav*，这是首个面向动态户外环境的基于场景图的语义导航框架。该系统集成了由大语言模型构建的多层级

具身图、检索增强推理与分层规划，从而实现具备实时适应能力的远程开放词汇导航。在仿真和真实场景中的大量实验表明，我们的方法能够有效提升户外导航的鲁棒性与效率。

讨论与局限性。尽管 CausalNav 在动态场景中表现有效，但在可扩展性、极端光照或天气条件下的鲁棒性以及长时程一致性方面仍面临局限。与 EmbodiedRAG [15] 和 NavRAG [16],等以往图记忆系统相比，CausalNav 通过双向 LLM–RAG推理和实时多层级图更新，将 RAG范式扩展至动态户外导航，实现了开放词汇理解和持续的空间语义适应。

未来工作。我们计划增强图压缩和记忆回忆机制，以提升可扩展性；探索多模态融合，以改善极端条件下的鲁棒性；并推动系统发展，实现长期自主探索和终身学习。

参考文献

[1] P. Yin、J. Jiao、S. Zhao、L. Xu、G. Huang、H. Choset、S. Scherer 和 J. Han，“通用地点识别综述：迈向现实世界自主性”，*IEEE*机器人学汇刊，第 1–20 页，2025年。[2] T. Guan、D. Kothandaraman、R. Chandra、A. J. Sathyamoorthy、K. Weer- akoon 和 D. Manocha，“ Ga-nav：面向非结构化户外环境中机器人导航的高效地形分割”，*IEEE*机器人与自动化快报，第 7 卷，第 3 期，第 8138–8145 页，2022年。[3] P. Yin、A. Abuduweili、S. Zhao、L. Xu、C. Liu 和 S. Scherer，“Bioslam：用于通用地点识别的仿生终身记忆系统”，*IEEE*机器人学汇刊，第 39 卷，第 6 期，第 4855–4874 页，2023 年。[4] F. Zhu 、 X. Liang 、 Y. Zhu 、 Q. Yu、X. Chang 和 X. Liang，“Soon：利用基于图的探索进行场景导向物体导航”，载于*2021*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），2021年，第 12 684–12 694 页。[5] S. Niijima、R. Umeyama、Y. Sasaki 和 H. Mizoguchi，“面向城市区域自主移动机器人导航的城市尺度网格-拓扑混合地图”，载于*2020*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），2020年，第 2065–2071 页。[6] D. Shah、B. Eysenbach、N. Rhinehart 和 S. Levine，“ 利用潜在目标模型实现开放世界导航的快速探索 ”，载于机器人学习会议。PMLR，2022年，第 674–684 页。[7] D. Shah 和 S. Levine，“ViKiNG：利用地理提示进行基于视觉的千米级导航”，载于《机器人：科学与系统》会议论文集，纽约市，纽约州，美国，2022年6月。[8] A. Sridhar、D. Shah、C. Glossop 和 S. Levine，“Nomad：用于导航与探索的目标掩码扩散策略”，载于*2024*年*IEEE*机器人与自动化国际会议（*ICRA*）。IEEE，2024年，第 63–70 页。[9] X. Liu、J. Li、Y. Jiang、N. Sujay、Z. Yang、J. Zhang、J. Abanes、J. Zhang 和 C. Feng，“Citywalker：从网络规模视频中学习具身城市导航”，载于计算机视觉与模式识别会议论文集，2025年，第 6875–6885 页。[10] N. Hughes、Y. Chang 和 L. Carlone，“ Hydra：用于三维场景图构建与优化的实时空间感知系统”，载于《机器人：科学与系统》，2022年。[11] K. Rana、J. Haviland、S. Garg、J. Abou-Chakra、I. Reid 和 N. Suender- hauf，“Sayplan：利用三维场景图为大语言模型提供依据，实现可扩展的机器人任务规划”，载于机器人学习会议。PMLR，2023年，第 23–72 页。[12] C. Huang、O. Mees、A. Zeng 和 W. Burgard，“用于机器人导航的视觉语言地图”，载于*2023*年*IEEE*机器人与自动化国际会议（*ICRA*）。IEEE，2023年，第 10 608–10 615 页。[13] C. Kassab、M. Mattamala、L. Zhang 和 M. Fallon，“语言扩展室内 SLAM（LExis）：用于实时视觉场景理解的多功能系统”，载于*2024*年*IEEE*机器人与自动化国际会议（*ICRA*）。IEEE，2024年，第 15 988–15 994 页。[14] Q. Xie、S. Y. Min、T. Zhang、K. Xu、A. Bajaj、R. Salakhutdinov、M. Johnson-Roberson 和 Y. Bisk，“具身检索增强生成：用于检索与生成的通用非参数具身记忆”，载于语言游戏化-NeurIPS 2024研讨会。[15] M. Booker、G. Byrd、B. Kemp、A. Schmidt 和 C. Rivera，“具身检索增强生成：用于高效、可扩展机器人任务规划的动态三维场景图检索”，arXiv 预印本 arXiv:2410.23968，2024年。[16] Z. Wang、Y. Zhu、G. H. Lee 和 Y. Fan，“Navrag：通过检索增强大语言模型为具身导航生成用户需求指令”，arXiv 预印本 arXiv:2502.11142，2025年。[17] C. Cadena、L. Carlone、H. Carrillo、Y. Latif、D. Scaramuzza、J. Neira、I. Reid 和 J. J. Leonard，“同时定位与建图的过去、现在与未来：迈向鲁棒感知时代”，IEEE机器人学汇刊，第 32 卷，第 6 期，第 1309–1332 页，2016年。[18] J. Hao、T. Yang、H. Tang、C. Bai、J. Liu、Z. Meng、P. Liu 和 Z. Wang，“深度强化学习中的探索：从单智能体到多智能体领域”，IEEE神经网络与学习系统汇刊，2023年。[19] D. Shah、A. Sridhar、N. Dashora、K. Stachowicz、K. Black、N. Hirose 和 S. Levine，“ViNT：用于视觉导航的基础模型”，载于机器人学习会议。PMLR，2023年，第 711–733 页。[20] J. Carvalho、A. T. Le、M. Baierl、D. Koert 和 J. Peters，“运动规划扩散：利用扩散模型学习和规划机器人运动”，载于2023年IEEE/RSJ智能机器人与系统国际会议（IROS）。IEEE，2023年，第 1916–1923 页。[21] D. Shah、B. Osi ́ński、S. Levine 等，“LM-Nav：利用语言、视觉与动作大型预训练模型实现机器人导航”，载于机器人学习会议。PMLR，2023年，第 492–504 页。[22] Y. Deng、J. Wang、J. Zhao、X. Tian、G. Chen、Y. Yang 和 Y. Yue，“OpenGraph：大规模室外环境中的开放词汇分层三维图表示”，IEEE机器人与自动化快报，2024年。[23] T. Cheng、L. Song、Y. Ge、W. Liu、X. Wang 和 Y. Shan，“Yolo-world：实时开放词汇目标检测”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2024年，第 16 901–16 911 页。[24] W. Xu、Y. Cai、D. He、J. Lin 和 F. Zhang，“Fast-lio2：快速直接激光雷达-惯性里程计”，IEEE机器人学汇刊，第 38 卷，第 4 期，第 2053–2073 页，2022年。[25] Y. Zhang、P. Sun、Y. Jiang、D. Yu、F. Weng、Z. Yuan、P. Luo、W. Liu 和 X. Wang，“ByteTrack：通过关联每个检测框实现多目标跟踪”，2022年。[26] D. Tsai、S. Worrall、M. Shan、A. Lohr 和 E. Nebot，“优化样本选择，以实现鲁棒的激光雷达-相机标定”，载于2021年IEEE智能交通系统国际会议（ITSC），2021年，第 2631–2638 页。[27] T. Yin、X. Zhou 和 P. Krahenbuhl，“基于中心的三维目标检测与跟踪”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2021年，第 11 784–11 793 页。[28] Y.-K. Lin、W.-C. Lin 和 C.-C. Wang，“用于激光雷达-惯性里程计的自运动估计与多目标跟踪异步状态估计”，载于2023年IEEE机器人与自动化国际会议（ICRA）。IEEE，2023年，第 10 616–10 622 页。[29] Y. Jia、T. Wang、F. Cao、X. Chen、S. Shao 和 L. Liu，“TRLO：一种高效激光雷达里程计，支持三维动态目标跟踪与移除”，IEEE仪器与测量汇刊，第 74 卷，10，2025 1pp. – 。[30] Z. Yan、X. Wu、Z. Jian、B. Lan 和 X. Wang，“ RH-Map：基于三维分区哈希地图结构的动态物体移除在线地图构建框架”，IEEE机器人与自动化快报，第 9 卷，第 2 期，第 1492–1499 页，2023年。[31] J. D. Gammell、S. S. Srinivasa 和 T. D. Barfoot，“Informed RRT*：通过直接采样可采纳椭球启发式实现聚焦的最优基于采样的路径规划”，载于2014年IEEE/RSJ智能机器人与系统国际会议，2014年，第 2997–3004 页。[32] J. Zeng、Z. Li 和 K. Sreenath，“利用离散时间控制屏障函数提升非线性模型预测控制的可行性与安全性”，载于2021年IEEE第60届决策与控制会议（CDC）。IEEE，2021年，第 6137–6144 页。[33] D. Shah、A. Sridhar、A. Bhorkar、N. Hirose 和 S. Levine，“GNM：驱动任意机器人的通用导航模型”，载于2023年IEEE机器人与自动化国际会议（ICRA），2023年，第 7226–7233 页。